AN EFFICIENT RESIDUAL ECHO SUPPRESSION FOR MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION BASED ON THE FREQUENCY-DOMAIN ADAPTIVE KALMAN FILTER

Jan Franzen, Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig,
Schleinitzstr. 22, 38106 Braunschweig, Germany
{j.franzen, t.fingscheidt}@tu-bs.de

ABSTRACT

Emerging use cases, such as keyword spotting while listening to FM radio, or participating in a teleconference utilizing the hands-free system in a vehicle, require the utilization of multi-channel acoustic echo cancellation (AEC). Addressing the typically remaining residual echo, it is common practice to apply a postfilter for residual echo suppression (RES) in a subsequent processing stage. In this paper we propose two RES approaches for the multi-channel frequency-domain adaptive Kalman filter, both being optimal under certain assumptions, extremely efficient and robust by exploiting a tight relation to the Kalman stepsize already available from the AEC.

Index Terms— residual echo suppression, multi-channel acoustic echo cancellation, frequency domain adaptive filter

1. INTRODUCTION

The problem of acoustic echo cancellation (AEC) has found wide interest over the past decades. Typically, an estimate of the echo path impulse response (IR) from loudspeaker to microphone is computed. Subsequently, the echo signal is estimated and then subtracted from the microphone signal. Facing this challenge, a wide range of solutions has been developed: time domain algorithms, as for example seen in [1], block-based or partitioned block-based processing algorithms, e.g., [2, 3, 4, 5, 6, 7], and sub-band processing algorithms, e.g., [8, 9, 10, 11, 12]. In the recent past, the frequency-domain adaptive filter (FDAF) using a state-space formulation according to Kalman filter theory [13, 14] enjoys increasing popularity. The approach shows fast convergence behavior, a robust double-talk performance without separate double-talk detection, and provides inherent stepsize control.

Although the aim of an AEC is to obtain an echo-free uplink signal, a residual echo still remains present in most cases. To address this, it is common practice to apply a postfilter for residual echo suppression (RES) [15, 16, 17, 18, 19]. In [13] a tight relation between the coefficients of a RES postfilter and the stepsize of the *single*-channel FDAF Kalman filter has been revealed. Based on this relation, the authors proposed a simple yet efficient RES approach for a mono AEC. In [7] this relation was as well validated for the *partitioned single*-channel FDAF Kalman filter.

Recently, more scenarios—such as the emerging use case of participating in a teleconference utilizing the hands-free system in a vehicle—cannot be sufficiently satisfied with a single-channel AEC but instead ask for stereo AEC (SAEC) or multi-channel solutions [20, 21, 22]. By estimating separate acoustic echo paths for two differing yet highly cross-correlated excitation signals, SAECs allow for two different acoustic transmission channels. In general, multi-channel algorithms furthermore have to face the so-called non-uniqueness problem [23]. For these scenarios either explicit decorrelation means [23, 24], or a robust FDAF solution employing the Kalman filter has successfully been employed [25, 26].

In this paper, we present an *efficient* RES postfilter for the *multi*-channel FDAF Kalman filter AEC. We expand the theoretical derivation of [13] from a single- to a multi-channel scenario showing also a tight relation between RES filter coefficients and the stepsize of the multi-channel FDAF Kalman filter, however, now including mixed terms. On the grounds of the theoretical derivation, we propose two *efficient* RES filter versions both being optimal under some given assumptions, one of them, however, with an additional assumption that interestingly leads to further increased robustness in the simulations.

The paper is structured as follows: In Section 2 a general time domain formulation of the optimal RES filter for the SAEC is derived. The efficient equivalent in the frequency domain is presented in Section 3, followed by an enhanced modified approach. Experimental evaluation and discussion of results are shown in Section 4. Section 5 provides conclusions.

2. TIME DOMAIN DERIVATION OF A STEREO ACOUSTIC ECHO CANCELLATION POSTFILTER

Our mathematical derivation in the time domain follows the easily comprehensible sample-wise convolutive notation employed by Enzner et al. in a related context [19]. In analogy to the mono case shown in [13], we extend the scenario to a *multi*-channel case and present in the following—without loss of generality—the stereo-channel case.

The microphone signal is received as

$$\begin{aligned} y(n) &= s(n) + d_1(n) + d_2(n) \\ &= s(n) + \mathbf{h}_1^T \mathbf{x}_1(n) + \mathbf{h}_2^T \mathbf{x}_2(n), \end{aligned} \tag{1}$$

with desired near-end signal $s(n)$ and two echo signals $d_j(n)$, $j \in \mathcal{I} = \{1, 2\}$. The echo signals $d_j(n)$ in turn depend on the K latest loudspeaker samples ($[\,]^T$ denotes the transpose)

$$\mathbf{x}_j(n) = [x_j(n), \ldots, x_j(n-(K-1))]^T, \quad j \in \mathcal{I}, \tag{2}$$

and the corresponding K coefficients of the IR from loudspeaker to microphone (for better readability written as if it were time-invariant) $\mathbf{h}_j = [h_{j,0}, \ldots, h_{j,K-1}]^T, j \in \mathcal{I}$.

The near-end signal $s(n)$ is treated as a stationary zero-mean random process. It is assumed as statistically independent from the loudspeaker signals $x_j(n)$, which are known and therefore

treated as deterministic. Statistically independent from both near-end and loudspeaker signals, the coefficient vectors $\mathbf{h}_j$ are seen—similar to [19]—as multivariate random variables with expectations $\bar{\mathbf{h}}_j = \mathcal{E}\{\mathbf{h}_j\}$, zero-mean unpredictable IR parts $\tilde{\mathbf{h}}_j = \mathbf{h}_j - \bar{\mathbf{h}}_j$, and covariance matrices $\mathbf{P}_{i,j} = \mathcal{E}\{\tilde{\mathbf{h}}_i \tilde{\mathbf{h}}_j^T\}$, $i, j \in \mathcal{I}$. Note that due to the multi-channel scenario the cross-channel covariances are new in this work.

Expanding (1) to vectorial notation containing the K latest samples and using (2) yields

$$\mathbf{y}(n) = \mathbf{s}(n) + \mathbf{X}_1^T(n)\mathbf{h}_1 + \mathbf{X}_2^T(n)\mathbf{h}_2 \tag{3}$$

with

$$\begin{aligned} \mathbf{y}(n) &= [y(n), \ldots, y(n-(K-1))]^T, \\ \mathbf{s}(n) &= [s(n), \ldots, s(n-(K-1))]^T, \\ \mathbf{X}_j(n) &= [\mathbf{x}_j(n), \ldots, \mathbf{x}_j(n-(K-1))], \quad j \in \mathcal{I}. \end{aligned} \tag{4}$$

The convolution operations of both SAEC and RES filter to obtain the estimated, ideally echo-free, near-end signal $\hat{s}(n)$ can be written in a similar manner:

$$\hat{s}(n) = \big(\mathbf{y}^T(n) - \hat{\mathbf{h}}_1^T \mathbf{X}_1(n) - \hat{\mathbf{h}}_2^T \mathbf{X}_2(n)\big)\mathbf{g}. \tag{5}$$

This equation includes two steps: First, subtracting estimated echo signals from the microphone signal using the estimated IRs $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$, $\hat{\mathbf{h}}_j = [\hat{h}_{j,0}, \ldots, \hat{h}_{j,K-1}]^T, j \in \mathcal{I}$, and afterwards the convolution with the RES filter IR $\mathbf{g} = [g_0, \ldots, g_{K-1}]^T$. Applying the minimum mean square error criterion (MMSE)

$$J = \mathcal{E}\big\{\big(s(n) - \hat{s}(n)\big)^2\big\} \rightarrow \text{min.}, \tag{6}$$

one can now compute the partial derivatives w.r.t. the estimated IRs $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$, and the RES filter IR $\mathbf{g}$. Details of the postfilter derivation can be found in the Appendix. The postfilter IR turns out to be

$$\boxed{\begin{aligned} \mathbf{g} = \big(&\mathbf{R}_s + \mathbf{X}_1^T(n)\mathbf{P}_{1,1}\mathbf{X}_1(n) + \mathbf{X}_1^T(n)\mathbf{P}_{1,2}\mathbf{X}_2(n) \\ &+ \mathbf{X}_2^T(n)\mathbf{P}_{2,1}\mathbf{X}_1(n) + \mathbf{X}_2^T(n)\mathbf{P}_{2,2}\mathbf{X}_2(n)\big)^{-1} \cdot \mathbf{r}_s. \end{aligned}} \tag{7}$$

3. NEW FREQUENCY DOMAIN POSTFILTER

To exploit the benefits of the previously derived optimal RES filter, a formulation in the frequency domain is required. Therefore, we bear in mind that for a *cyclic* matrix $\mathbf{A}_{K\times K}$ the multiplication with a vector $\mathbf{b}_{1\times K}$ equals K result samples of a cyclic convolution of the vectors $\mathbf{a}$ and $\mathbf{b}$, where $\mathbf{a}$ is the first column of the matrix $\mathbf{A}$. Hence, the K-point Fourier transform of (7) with bin index k can be found to provide the coefficients (see also [27])

$$\boxed{\begin{aligned} G(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_1(\ell,k)\Phi_{1,2}(\ell,k)X_2^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{ss}(\ell,k), \end{aligned}} \tag{8}$$

incorporating the approximation that correlation and covariance matrices are now calculated as *cyclic* and based on signal *frames* with index ℓ. Here, $\Phi_{ss}(\ell,k)$ is the power spectral density of the near-end signal, and $\big(\Phi_{i,j}(\ell,k)\big)\big|_{k=0}^{k=K-1} \approx \text{DFT}\big\{\text{diag}\{\mathbf{P}_{i,j}\}\big\}$ are the so-called residual echo power transfer functions (cf. [13, p. 1143]), while diag{} extracts the main diagonal from its matrix argument.

In preparation of the efficient RES filter formulations later on, we now remind the reader of the SAEC stepsizes for the FDAF Kalman filter as presented in [26, eq. (10)]:

$$\begin{aligned} \mu_{i,j}(\ell,k) = \Big(&\Psi_{ss}(\ell,k) + \tfrac{R}{K}\big(X_1(\ell,k)P_{1,1}^+(\ell,k)X_1^*(\ell,k) \\ &+ X_1(\ell,k)P_{1,2}^+(\ell,k)X_2^*(\ell,k) \\ &+ X_2(\ell,k)P_{2,1}^+(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)P_{2,2}^+(\ell,k)X_2^*(\ell,k)\big)\Big)^{-1} \cdot \tfrac{R}{K}P_{i,j}^+(\ell,k). \end{aligned} \tag{9}$$

Here, R is the frame-shift, $\big(P_{i,j}^+(\ell,k)\big)\big|_{k=0}^{k=K-1} = \text{diag}\{\mathbf{P}_{i,j}^+(\ell)\}$ with the predicted $K \times K$ diagonal state error covariance matrices $\mathbf{P}_{i,j}^+(\ell)$, and $\big(\Psi_{ss}(\ell,k)\big)\big|_{k=0}^{k=K-1} = \text{diag}\{\mathbf{\Psi}_{ss}(\ell)\}$ with the $K \times K$ diagonal measurement noise covariance matrix $\mathbf{\Psi}_{ss}(\ell)$. With the approximations $\Psi_{ss}(\ell,k) \approx R \cdot \Phi_{ss}(\ell,k)$ and $P_{i,j}^+(\ell,k) \approx K \cdot \Phi_{i,j}(\ell,k)$ this can be rewritten as:

$$\begin{aligned} \mu_{i,j}(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_1(\ell,k)\Phi_{1,2}(\ell,k)X_2^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{i,j}(\ell,k). \end{aligned} \tag{10}$$

3.1. Efficient Optimal Postfilter Based on Stepsize

Comparing (8) and (10) reveals that the relation Enzner et al. proposed for the single-channel case [13, eq. (77)] can be found *with mixed terms* in the multi-channel case. Different to [27], the optimal RES filter coefficients are easily and efficiently obtained from the stepsizes as

$$\boxed{G_{\text{PFI}}(\ell,k) = 1 - \Big(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} X_i(\ell,k)\mu_{i,j}(\ell,k)X_j^*(\ell,k)\Big).} \tag{11}$$

The structure of this equation further reveals that the extension from stereo to even more channels is intuitively obvious by extending the definition of the channel index set $\mathcal{I}$.

3.2. Efficient Optimal Postfilter Based on Stepsize with Additional Assumption

The derivation of the RES filter in Section 3.1 allows to model some statistical dependence between the unpredictable residual parts of the two IRs, i.e.,

$$\mathbf{P}_{i,j} = \mathcal{E}\{\tilde{\mathbf{h}}_i \tilde{\mathbf{h}}_j^T\} \neq 0, \quad i \neq j, \quad i, j \in \mathcal{I}. \tag{12}$$

Since the unpredictable IR parts $\tilde{\mathbf{h}}_j$ reveal typical properties of random processes, it may be reasonable to additionally assume that $\mathbf{P}_{1,2} = \mathbf{P}_{2,1} = \mathbf{0}$ with $\mathbf{0}$ being the $K \times K$ zero matrix. In consequence, the mixed terms in (7) vanish, and hence (8) becomes (see also [27])

$$\begin{aligned} G(\ell,k)) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{ss}(\ell,k). \end{aligned} \tag{13}$$

Applying this assumption also to the definition of the stepsize (10) yields a formulation that can now be used for the postfilter[1]:

$$\begin{aligned} \mu_j(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{j,j}(\ell,k). \end{aligned} \tag{14}$$

Analog to (11) and different to [27], we now easily find the optimal postfilter coefficients with a *zero-correlation* assumption as

[1] Note that (14) can be efficiently computed from parts of the additions in (10), the latter still being the stepsize to be used in the AEC.

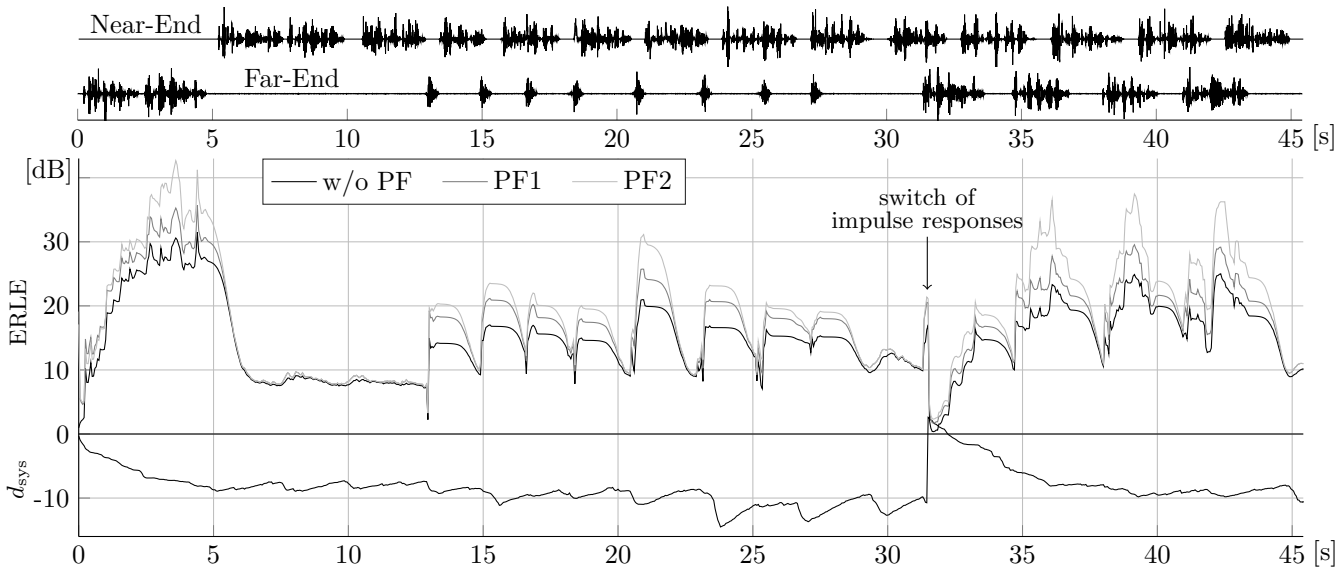

Fig. 1. Speech waveforms (top) and results for the baseline without postfilter and the two proposed postfilter approaches. ERLE (center curves) and system distance (bottom curves) over time: single-talk (0-10 s), single-talk with barge-ins (10-30 s), double-talk (30-45 s).

$$G_{\text{PF2}}(\ell,k) = 1 - \big(\sum_{j\in\mathcal{I}} X_j(\ell,k)\mu_j(\ell,k)X_j^*(\ell,k)\big). \tag{15}$$

Note that—as before—all used terms are directly available from the SAEC algorithm, and the extension from stereo to more channels is again straightforward.

4. EXPERIMENTAL VALIDATION

For the experimental validation we are using the FDAF SAEC in a robust hands-free configuration similar to [26]. At a sampling frequency of 16 kHz, the DFT size is set to $K = 1024$. The FDAF parameters are: forgetting factor $A = 0.998$, overestimation factor $\lambda = 1.5$ and $\mathbf{\Psi}_{ss}$ smoothing factor $\beta = 0.5$ (see [26, eq. 12]). Ensuring linear convolution in the echo cancellation path, $K - R$ coefficients remain to cover the echo path IRs. This equals a length of 48 ms. The postfilter is applied in a subsequent step: In each frame the most recent $2R$ samples of the AEC output are multiplied with a square root Hann window and zero-padded to be subject to a K-point DFT. Now the K postfilter coefficients are applied, and the result is again subject to a K-point IDFT. The first $2R$ output samples are multiplied with the second square root Hann window, and finally combined with the previous frame's output according to the window overlap. We furthermore apply some well-established constraints to the postfilter coefficients, in order to avoid artifacts. These imply smoothing over time by

$$G_{\text{final}}(\ell,k) = 0.5\cdot G_{\{\text{PF1, PF2}\}}(\ell,k) + 0.5\cdot G_{\text{final}}(\ell-1,k), \tag{16}$$

and a gain limitation to the value range $[0.05, \ldots, 1]$.

ITU-T Recommendation P.501 [28, Secs. 7.3.5, 7.3.7] test signals are used for simulating a challenging double-talk scenario and composed as follows: For the far-end, speech signal $s'(n)$ is made up of the 10 s *short conditioning sequence I* followed by the ca. 35 s *double-talk sequence*, both with at a level of -26 dBov. It is convolved with randomly generated far-end IRs $h'_j(n)$ with exponential energy decay and a car-typical reverberation time of $T_{60} = 50$ ms. The generated IRs are cut off after 50 ms. Uncorrelated white Gaussian noise is added with a level of -66 dBov as sensor noise $n'_i(n)$, thereby yielding the loudspeaker signals $x_j(n)$. In the same way the near-end IRs $h_j(n)$ are created and convolved with the loudspeaker signals to obtain the two echo signals $d_j(n)$. Then *short conditioning sequence II* followed by the *single-talk sequence* are added as near-end speech $s(n)$ at a level of -26 dBov, and *in-car noise* is added as near-end noise $n(n)$ at an SNR of 15 dB to finally obtain the microphone signal $y(n)$. Additionally, the near-end impulse responses are switched during double-talk after 31.5 s.

Figure 1 shows the speech waveforms on the top and the experimental results of the postfilter performance below. The center curves show the echo return loss enhancement (ERLE)

$$\text{ERLE}(n) = 10\log\left(\frac{d^2(n)}{\big(d(n)-\hat{d}(n)\big)^2}\right), \tag{17}$$

with $d(n) = d_1(n) + d_2(n)$ and $\hat{d}(n) = \hat{d}_1(n) + \hat{d}_2(n)$. It is computed as done in [14]. The lower curve depicts the course of the system distance ($\mathcal{I} = \{1, 2\}$)

$$d_{\text{sys}}(n) = 10\log\left(\frac{\sum_{j\in\mathcal{I}} ||\mathbf{h}_j(n) - \hat{\mathbf{h}}_j(n)||^2}{\sum_{j\in\mathcal{I}} ||\mathbf{h}_j(n)||^2}\right). \tag{18}$$

The baseline without any RES postfilter reaches up to 30 dB ERLE in single-talk, ca. 15 dB at far-end barge-ins, and around 20 dB during harsh double-talk. The system distance is the same for all approaches and shows the fast convergence of the AEC algorithm, even after the switch of impulse responses during double-talk. Aside from the two convergence periods, the system distance reaches about -10 dB. Applying the newly derived RES postfilter G_{PF1} (11) directly increases the ERLE in all speech sections. The highest increases of up to 5 dB can be seen at single-talk and later in the double-talk sequence. The second proposed RES postfilter G_{PF2} (15) leads to a further consistent increase of ERLE during all sections also being clearly higher than the first approach. During single-talk ERLE values of more than 40 dB are reached. While double-talk is present, up to 13 dB of additional echo suppression are achieved by the postfilter G_{PF2}.

Table 1 additionally provides PESQ [29] and mean ERLE values for the experiments with an SNR of 0 dB and 15 dB. Since the system

	SNR 0 dB			SNR 15 dB		
	w/o PF	**PF1**	**PF2**	**w/o PF**	**PF1**	**PF2**
MOS($\hat{s}_{\text{PF}}$)	1.31	1.32	1.34	2.21	2.36	2.47
MOS($\tilde{s}_{\text{PF}}$)	4.64	4.59	4.47	4.64	4.59	4.49
ERLE	12.27	14.18	15.73	14.88	17.05	18.83

Table 1. MOS($\tilde{s}_{\text{PF}}$), MOS($\hat{s}_{\text{PF}}$), both LQO, and mean ERLE [dB] for noisy speech scenarios at SNRs of 0 dB and 15 dB.

distance does not change with different postfilters, it is not explicitly listed in the table.

For an SNR of 15 dB and including both (initial) convergence stages, a *mean* system distance of -8.4 dB is reached. The PESQ MOS listening quality objective (LQO) computed on the enhanced signal $\hat{s}_{\text{PF}}(n)$ with $s(n)$ as reference shows an improvement of 0.15 points with postfilter 1 (PF1) and another 0.11 points with postfilter 2 (PF2). Since PESQ has not been validated for artifacts caused by noise reduction techniques—which the RES is in this case—, we also provide the MOS LQO computed only on the processed clean speech component $\tilde{s}_{\text{PF}}(n)$ [30, sec. 8] obtained by using [31]. The speech *component* quality MOS($\tilde{s}_{\text{PF}}$) of both postfilters stays above a comfortable 4.4 MOS points indicating almost no distortion. The mean ERLE, however, is increased by 2.17 dB (PF1) and by even 3.95 dB (PF2), respectively. This is in line with our subjective listening impressions: While the far-end speech is being suppressed, none of the postfilter approaches practically affects the near-end speech at all.

For an SNR of 0 dB, the mean system distance is -6 dB. PESQ and mean ERLE here reveal a behavior analog to the 15 dB SNR case, thus leaving G_{PF2} with the best residual echo suppression performance. Although we took a further assumption of $\Phi_{i,j}(\ell,k)=0$ for $i\neq j$ to the derivation of G_{PF2}, it seems that this results in better overall performance than to include the estimates of $\Phi_{i,j}(\ell,k), i\neq j$ from the AEC as well for the postfilter.

Similar results and tendencies can be observed using the so-called black box approach to obtain $\tilde{s}_{\text{PF}}(n)$ and $\tilde{d}_{\text{PF}}(n)$ according to ITU-T Recommendation P.1110 [30, sec. 8] and [32, 33].

5. CONCLUSIONS

In this paper, we proposed two *efficient* residual echo suppression (RES) approaches for the *multi*-channel frequency-domain adaptive Kalman filter AEC. Based on a theoretical derivation in the time domain, a tight relation between RES filter coefficients and the stepsizes of the multi-channel AEC has been shown. Using this relation, the coefficients are remarkably efficient to obtain. The experimental results show a robust performance with up to 5 dB additional RES for the first approach, and up to 13 dB of additional echo suppression for the second approach.

APPENDIX

Remembering the previously introduced statistical assumptions, using (5), and assuming $\mathbf{g}$ and $\hat{\mathbf{h}}_j$ to be deterministic yields

$$\begin{aligned}\frac{\partial J}{\partial \hat{\mathbf{h}}_1} &= -2\mathcal{E}\big\{\big(s(n)-\hat{s}(n)\big)\frac{\partial \hat{s}(n)}{\partial \hat{\mathbf{h}}_1}\big\}\\ &= -2\mathcal{E}\big\{\big(s(n)-\hat{s}(n)\big)\big\}\cdot\frac{\partial \hat{s}(n)}{\partial \hat{\mathbf{h}}_1}\\ &= -2\mathcal{E}\big\{\hat{s}(n)\big\}\mathbf{X}_1(n)\mathbf{g}\\ &= -2\big(\mathcal{E}\big\{\mathbf{y}^T(n)\big\}-\hat{\mathbf{h}}_1^T\mathbf{X}_1(n)\\ &\qquad -\hat{\mathbf{h}}_2^T\mathbf{X}_2(n)\big)\mathbf{g}\cdot\mathbf{X}_1(n)\mathbf{g}.\end{aligned} \tag{19}$$

Taking the expectation of (3) in transposed form yields

$$\mathcal{E}\big\{\mathbf{y}^T(n)\big\}=\mathcal{E}\big\{\mathbf{h}_1^T\big\}\mathbf{X}_1(n)+\mathcal{E}\big\{\mathbf{h}_2^T\big\}\mathbf{X}_2(n). \tag{20}$$

We find that by picking the estimated IRs in (19) as the deterministic IR part (i.e., $\hat{\mathbf{h}}_j^T=\mathcal{E}\left\{\mathbf{h}_j^T\right\}, j\in\mathcal{I}$), the term $\mathcal{E}\left\{\hat{s}(n)\right\}=0$ and thereby equation (19) equals zero. This is in line with the mono case result of [19, eq. (23)], and the same holds for the derivative w.r.t. $\hat{\mathbf{h}}_2$.

Using (5), and $\mathcal{E}\left\{\hat{s}(n)\right\}=0$ as stated above, the derivative w.r.t. the RES filter IR yields

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{g}} &= -2\mathcal{E}\big\{\big(s(n)-\hat{s}(n)\big)\frac{\partial \hat{s}(n)}{\partial \mathbf{g}}\big\}\\ &= -2\mathcal{E}\big\{\big(s(n)-\hat{s}(n)\big)\big(\mathbf{y}(n)\\ &\qquad -\mathbf{X}_1^T(n)\hat{\mathbf{h}}_1-\mathbf{X}_2^T(n)\hat{\mathbf{h}}_2\big)\big\}\\ &= -2\mathcal{E}\big\{\mathbf{y}(n)\big(s(n)-\hat{s}(n)\big)\big\}\\ &\qquad -2\mathcal{E}\big\{\hat{s}(n)\big\}\big(\mathbf{X}_1^T(n)\hat{\mathbf{h}}_1+\mathbf{X}_2^T(n)\hat{\mathbf{h}}_2\big)\\ &= -2\mathcal{E}\big\{\mathbf{y}(n)\big(s(n)-\hat{s}(n)\big)\big\}.\end{aligned} \tag{21}$$

Using (5) and (3), we obtain for the above expectation

$$\begin{aligned}&\mathcal{E}\big\{\mathbf{y}(n)\big(s(n)-\hat{s}(n)\big)\big\}\\ &=\mathcal{E}\big\{\mathbf{y}(n)\cdot\\ &\quad\Big(s(n)-\big(\mathbf{y}^T(n)-\hat{\mathbf{h}}_1^T\mathbf{X}_1(n)-\hat{\mathbf{h}}_2^T\mathbf{X}_2(n)\big)\mathbf{g}\Big)\big\}\\ &=\mathcal{E}\big\{\big(\mathbf{s}(n)+\mathbf{X}_1^T(n)\mathbf{h}_1+\mathbf{X}_2^T(n)\mathbf{h}_2\big)\cdot\\ &\quad\Big(s(n)-\big(\mathbf{s}^T(n)+\mathbf{h}_1^T\mathbf{X}_1(n)+\mathbf{h}_2^T\mathbf{X}_2(n)\\ &\qquad -\hat{\mathbf{h}}_1^T\mathbf{X}_1(n)-\hat{\mathbf{h}}_2^T\mathbf{X}_2(n)\big)\mathbf{g}\Big)\big\}.\end{aligned} \tag{22}$$

Inserting $\mathbf{h}_j=\bar{\mathbf{h}}_j+\tilde{\mathbf{h}}_j$ and $\hat{\mathbf{h}}_j=\bar{\mathbf{h}}_j, j\in\mathcal{I}$, we obtain

$$\begin{aligned}\cdots &=\mathcal{E}\big\{\Big(\mathbf{s}(n)+\mathbf{X}_1^T(n)\big(\bar{\mathbf{h}}_1+\tilde{\mathbf{h}}_1\big)+\mathbf{X}_2^T(n)\big(\bar{\mathbf{h}}_2+\tilde{\mathbf{h}}_2\big)\Big)\cdot\\ &\qquad\Big(s(n)-\big(\mathbf{s}^T(n)+\tilde{\mathbf{h}}_1^T\mathbf{X}_1(n)+\tilde{\mathbf{h}}_2^T\mathbf{X}_2(n)\big)\mathbf{g}\Big)\big\}\\ &=\mathcal{E}\big\{\mathbf{s}(n)s(n)\big\}-\mathcal{E}\big\{\mathbf{s}(n)\mathbf{s}^T(n)\mathbf{g}\big\}\\ &\quad-\mathcal{E}\big\{\mathbf{X}_1^T(n)\tilde{\mathbf{h}}_1\tilde{\mathbf{h}}_1^T\mathbf{X}_1(n)\mathbf{g}\big\}\\ &\quad-\mathcal{E}\big\{\mathbf{X}_1^T(n)\tilde{\mathbf{h}}_1\tilde{\mathbf{h}}_2^T\mathbf{X}_2(n)\mathbf{g}\big\}\\ &\quad-\mathcal{E}\big\{\mathbf{X}_2^T(n)\tilde{\mathbf{h}}_2\tilde{\mathbf{h}}_1^T\mathbf{X}_1(n)\mathbf{g}\big\}\\ &\quad-\mathcal{E}\big\{\mathbf{X}_2^T(n)\tilde{\mathbf{h}}_2\tilde{\mathbf{h}}_2^T\mathbf{X}_2(n)\mathbf{g}\big\}\\ &=\mathbf{r}_s-\mathbf{R}_s\mathbf{g}\\ &\quad-\mathbf{X}_1^T(n)\mathbf{P}_{1,1}\mathbf{X}_1(n)\mathbf{g}-\mathbf{X}_1^T(n)\mathbf{P}_{1,2}\mathbf{X}_2(n)\mathbf{g}\\ &\quad-\mathbf{X}_2^T(n)\mathbf{P}_{2,1}\mathbf{X}_1(n)\mathbf{g}-\mathbf{X}_2^T(n)\mathbf{P}_{2,2}\mathbf{X}_2(n)\mathbf{g}\\ &\stackrel{!}{=}0,\end{aligned} \tag{23}$$

where $\mathbf{R}_s$ denotes the autocorrelation matrix of $s(n)$ and $\mathbf{r}_s$ is the matrix's first column. For the optimal RES filter coefficients we obtain a representation that corresponds to Enzner et al.'s formulation of a *generalized Wiener filter* (cf. [19, eq. 24] or [13, eq. 11]) and now further includes stereo-channel (or multi-channel) related terms as shown in (7).

6. REFERENCES

[1] H. Shin, A. H. Sayed, and W. Song, "Variable Step-Size NLMS and Affine Projection Algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, Feb. 2004.

[2] J. Lee and C. Un, "Block Realization of Multirate Adaptive Digital Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 105–117, Feb. 1986.

[3] S. Malik and G. Enzner, "State-Space Frequency-Domain Adaptive Filtering for Nonlinear Acoustic Echo Cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065–2079, Sept. 2012.

[4] E. Moulines, O. Ait Amrane, and Y. Grenier, "The Generalized Multidelay Adaptive Filter: Structure and Convergence Analysis," *IEEE Transactions on Signal Processing*, vol. 43, no. 1, pp. 14–28, Jan. 1995.

[5] Bernhard H. Nitsch, "A Frequency-Selective Stepfactor Control for an Adaptive Filter Algorithm Working in the Frequency Domain," *Signal Processing (Elsevier)*, vol. 80, no. 9, pp. 1733–1745, Sept. 2000.

[6] J. S. Soo and K. K. Pang, "Multidelay Block Frequency Domain Adaptive Filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, Feb. 1990.

[7] F. Kuech, E. Mabande, and G. Enzner, "State-Space Architecture of the Partitioned-Block-Based Acoustic Echo Controller," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 1295–1299.

[8] A. Gilloire and M. Vetterli, "Adaptive Filtering in Subbands with Critical Sampling: Analysis, Experiments, and Application to Acoustic Echo Cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, Aug. 1992.

[9] J. Chen, H. Bes, J. Vandewalle, and P. Janssens, "A New Structure for Sub-Band Acoustic Echo Canceler," in *Proc. of ICASSP*, New York, NY, USA, Apr. 1988, pp. 2574–2577.

[10] G. Schmidt, "Acoustic Echo Control in Subbands - An Application of Multirate Systems," in *Proc. of EUSIPCO*, Rhodes, Greece, Sept. 1998, pp. 1961–1964.

[11] W. Kellermann, "Analysis and Design of Multirate Systems for Cancellation of Acoustical Echoes," in *Proc. of ICASSP*, New York, NY, USA, Apr. 1988, pp. 2570–2573.

[12] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Hands-free System with Low-Delay Subband Acoustic Echo Control and Noise Reduction," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 1521–1524.

[13] G. Enzner and P. Vary, "Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Hands-Free Telephones," *Signal Processing (Elsevier)*, vol. 86, no. 6, pp. 1140–1156, June 2006.

[14] M.-A. Jung and T. Fingscheidt, "A Shadow Filter Approach to a Wideband FDAF-Based Automotive Handsfree System," in *5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, Sept. 2011.

[15] R. Martin and J. Altenhoner, "Coupled Adaptive Filters for Acoustic Echo Control and Noise Reduction," in *Proc. of ICASSP*, Detroit, MI, USA, May 1995, vol. 5, pp. 3043–3046.

[16] S. Goetze, K. D. Kammeyer, M. Kallinger, and A. Mertins, "Enhanced Partitioned Stereo Residual Echo Estimation," in *Fortieth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Oct. 2006, pp. 1326–1330.

[17] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral Feature-Based Nonlinear Residual Echo Suppression," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2013.

[18] J. Wung, T. S. Wada, B. H. Juang, B. Lee, T. Kalker, and R. W. Schafer, "A System Approach to Residual Echo Suppression in Robust Hands-Free Teleconferencing," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011, pp. 445–448.

[19] G. Enzner, R. Martin, and P. Vary, "Uncertainty Modeling in Acoustic Echo Control," in *48th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2014, pp. 1633–1638.

[20] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A Hybrid Mono/Stereo Acoustic Echo Canceler," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 468–475, Sept. 1998.

[21] H. Buchner and W. Kellermann, "Acoustic Echo Cancellation for Two and More Reproduction Channels," in *Proc. of International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, Sept. 2001, pp. 99–102.

[22] C. Stanciu, C. Paleologu, J. Benesty, S. Ciochina, and F. Albu, "Variable-Forgetting Factor RLS for Stereophonic Acoustic Echo Cancellation with Widely Linear Model," in *Proc. of EUSIPCO*, Bucharest, Romania, Aug. 2012, pp. 1960–1964.

[23] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic Acoustic Echo Cancellation - An Overview of the Fundamental Problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[24] J. Herre, H. Buchner, and W. Kellermann, "Acoustic Echo Cancellation for Surround Sound using Perceptually Motivated Convergence Enhancement," in *Proc. of ICASSP*, Honolulu, HI, USA, Apr. 2007, pp. I–17–I–20.

[25] S. Malik and G. Enzner, "Recursive Bayesian Control of Multichannel Acoustic Echo Cancellation," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 619–622, Nov. 2011.

[26] M. A. Jung, S. Elshamy, and T. Fingscheidt, "An Automotive Wideband Stereo Acoustic Echo Canceler using Frequency-Domain Adaptive Filtering," in *Proc. of EUSIPCO*, Lisbon, Portugal, Sept. 2014, pp. 1452–1456.

[27] S. Malik, *Bayesian Learning of Linear and Nonlinear Acoustic System Models in Hands-free Communication*, Ph.D. thesis, Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany, Oct. 2012.

[28] "ITU-T Recommendation P.501, Test signals for use in telephonometry," ITU, Jan. 2012.

[29] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Nov. 2007.

[30] "ITU-T Recommendation P.1110, Wideband Hands-Free Communication in Motor Vehicles," ITU, Mar. 2017.

[31] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[32] T. Fingscheidt and S. Suhadi, "Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo," in *Proc. of INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 818–821.

[33] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 273–276.